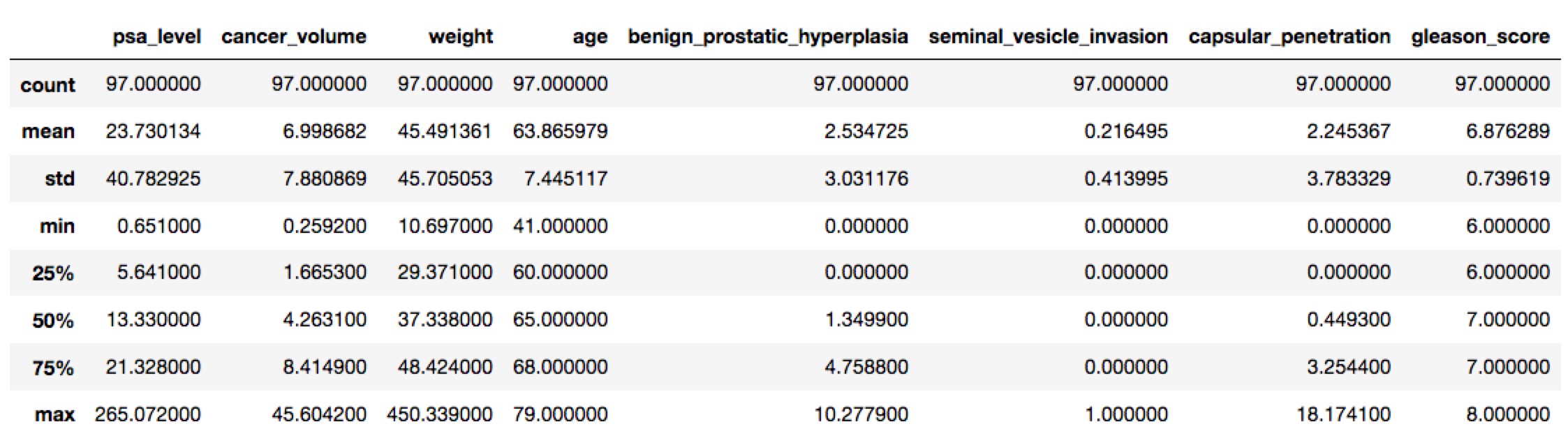
**Question 1:**

Descriptive statistics for the original dataset:



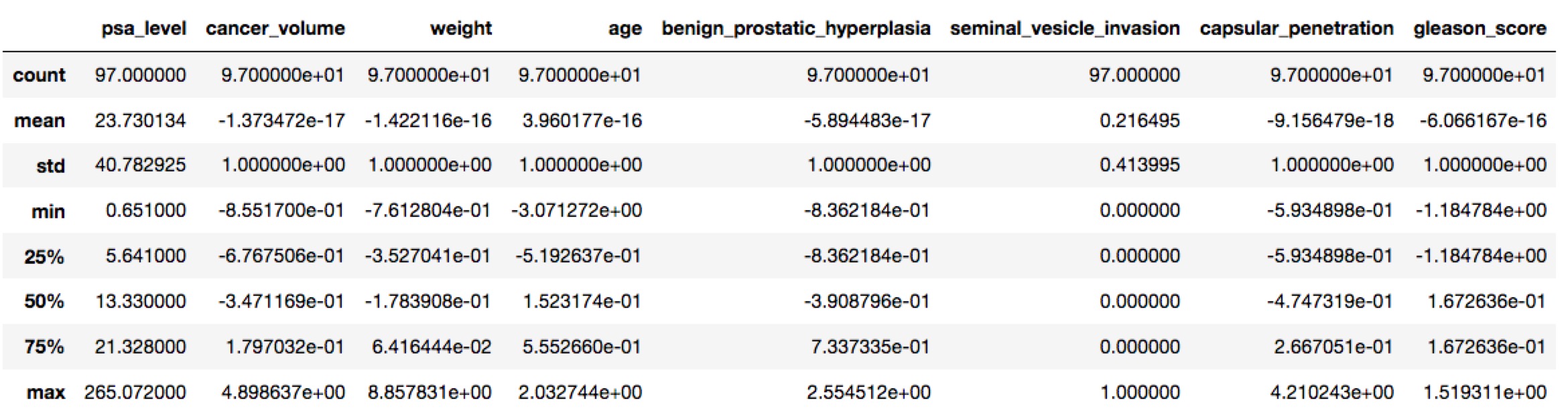
Based on the description of the data, we can see that “seminal\_vesicle\_invasion” is a categorical variable, and others can be considered numerical variables.

So, for linear regression model, assume the full model is:



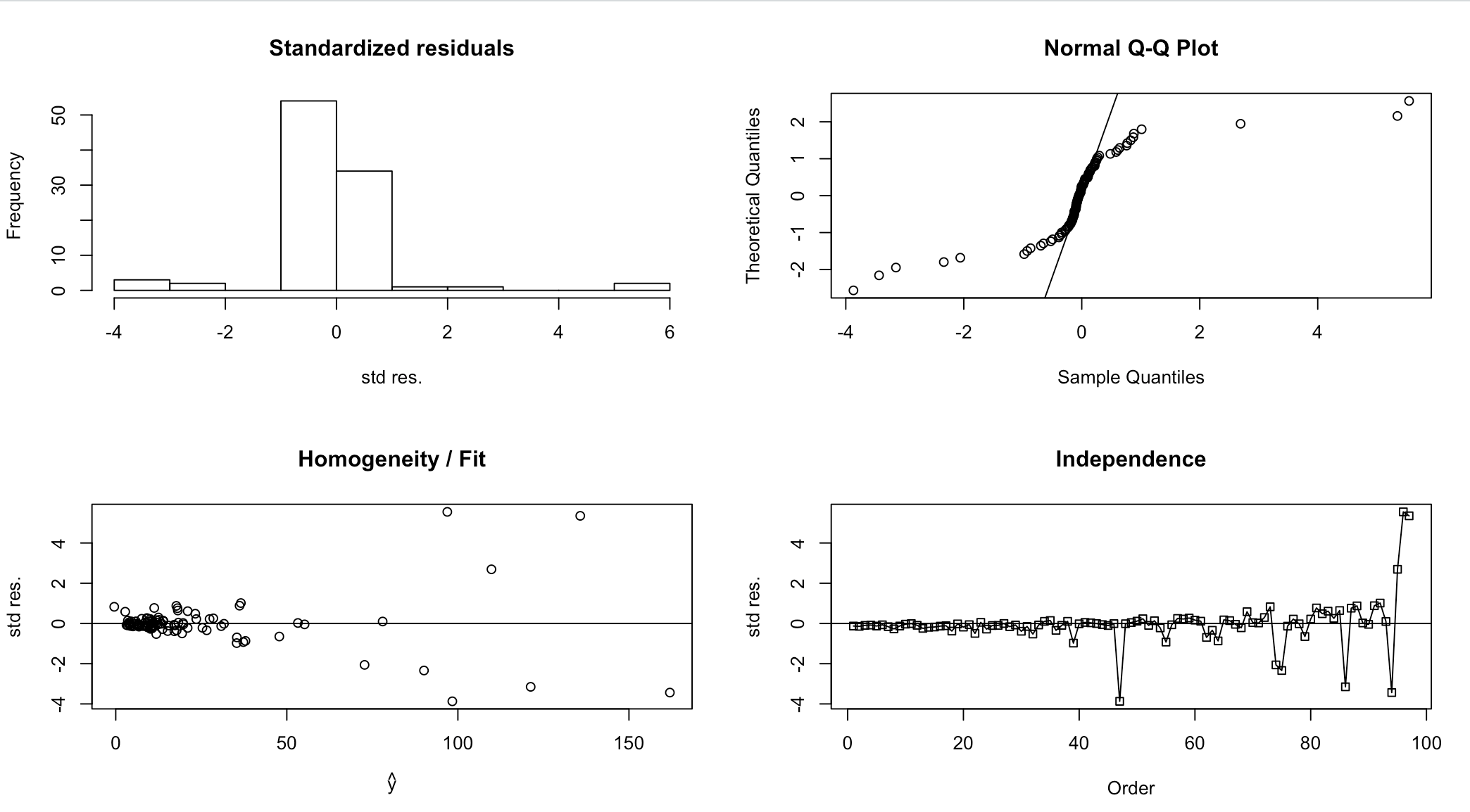
(y is psa\_level; x1 is cancer\_volumn; x2 is weight; x3 is age; x4 is benign\_prostatic\_hyperplasia; x5 is seminal\_vasicle\_invasion; x6 is capsular\_penetration; x7 is gleason\_score)

Based on the descriptive statistic above, the numerical predictors are obviously in different scales, which would lead to uninterpretable coefficients (We couldn’t see the impact of a predictor towards the response based on its coefficient in this sense). So, I transferred them to 0 mean, 1 standard deviation, and did the descriptive statistics again:



Now all the mean values of numerical predictors are 0, and all the standard deviations of them are 1.

Fit the model and check the assumption:

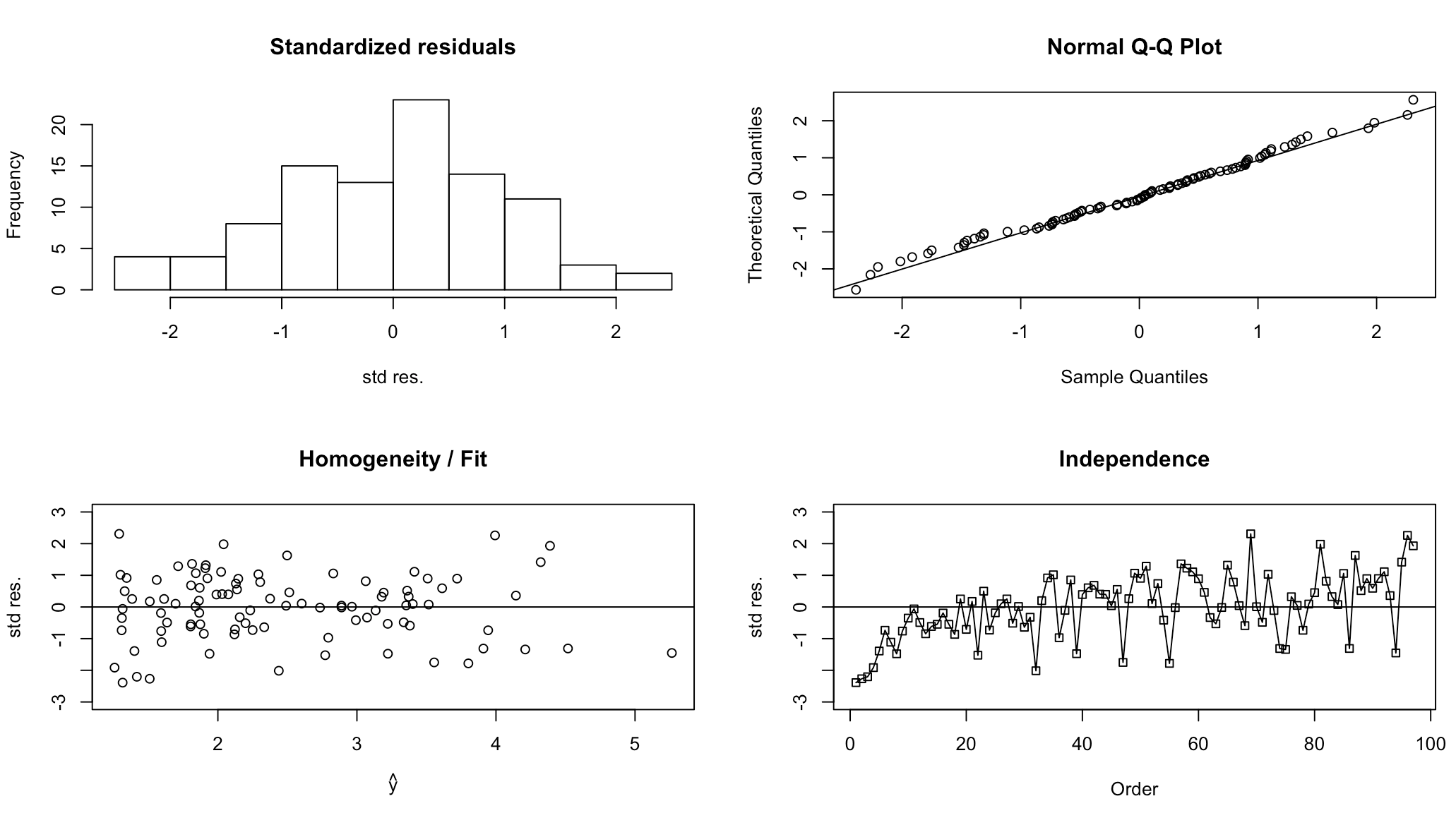


This is obviously not eligible for any further analysis (histogram and Q-Q plot show that it did not meet the normality requirement; and the spread of residuals were not constant; further, the pattern of independence check was almost discernable).

So, convert y to log(y):

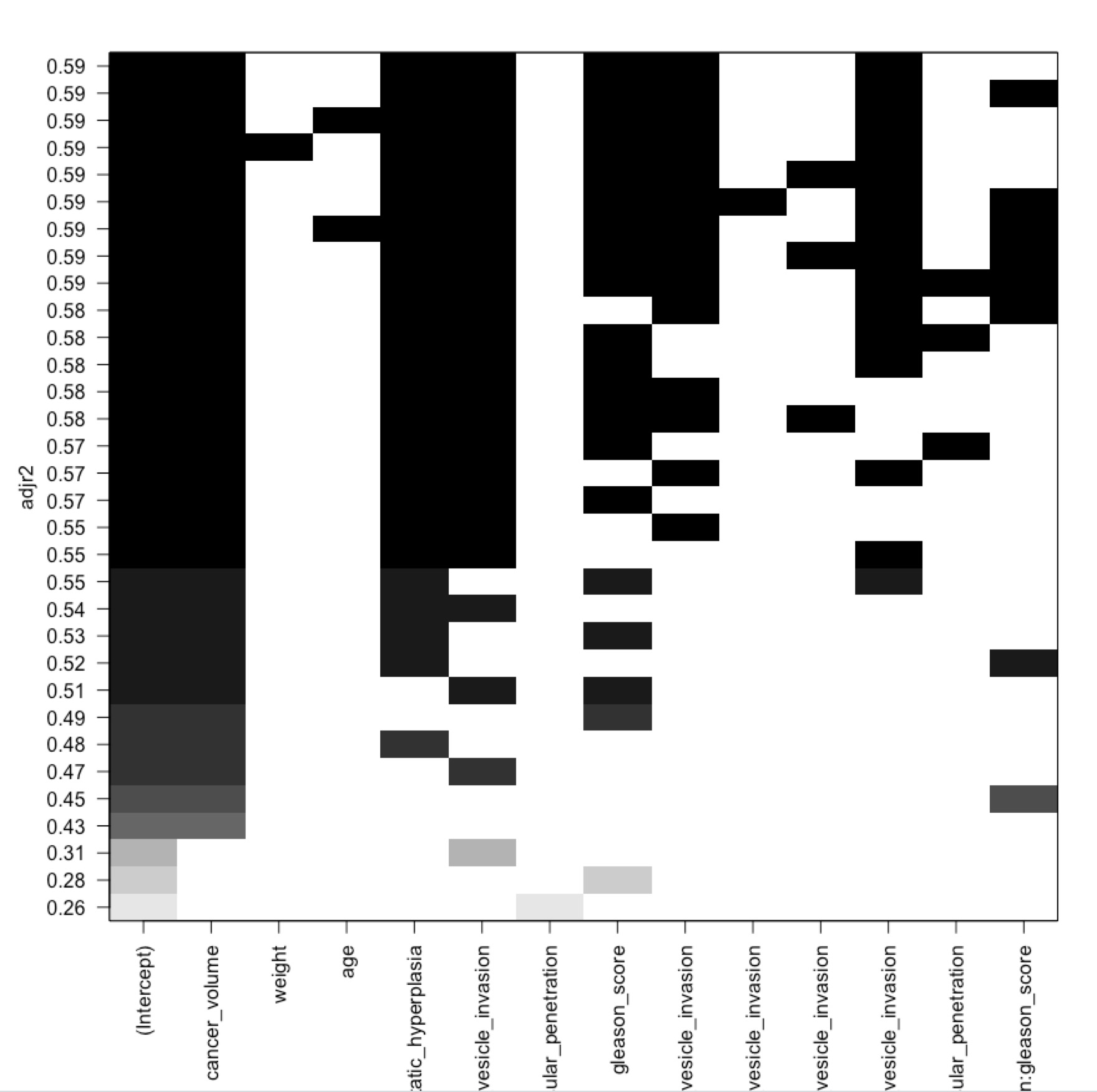
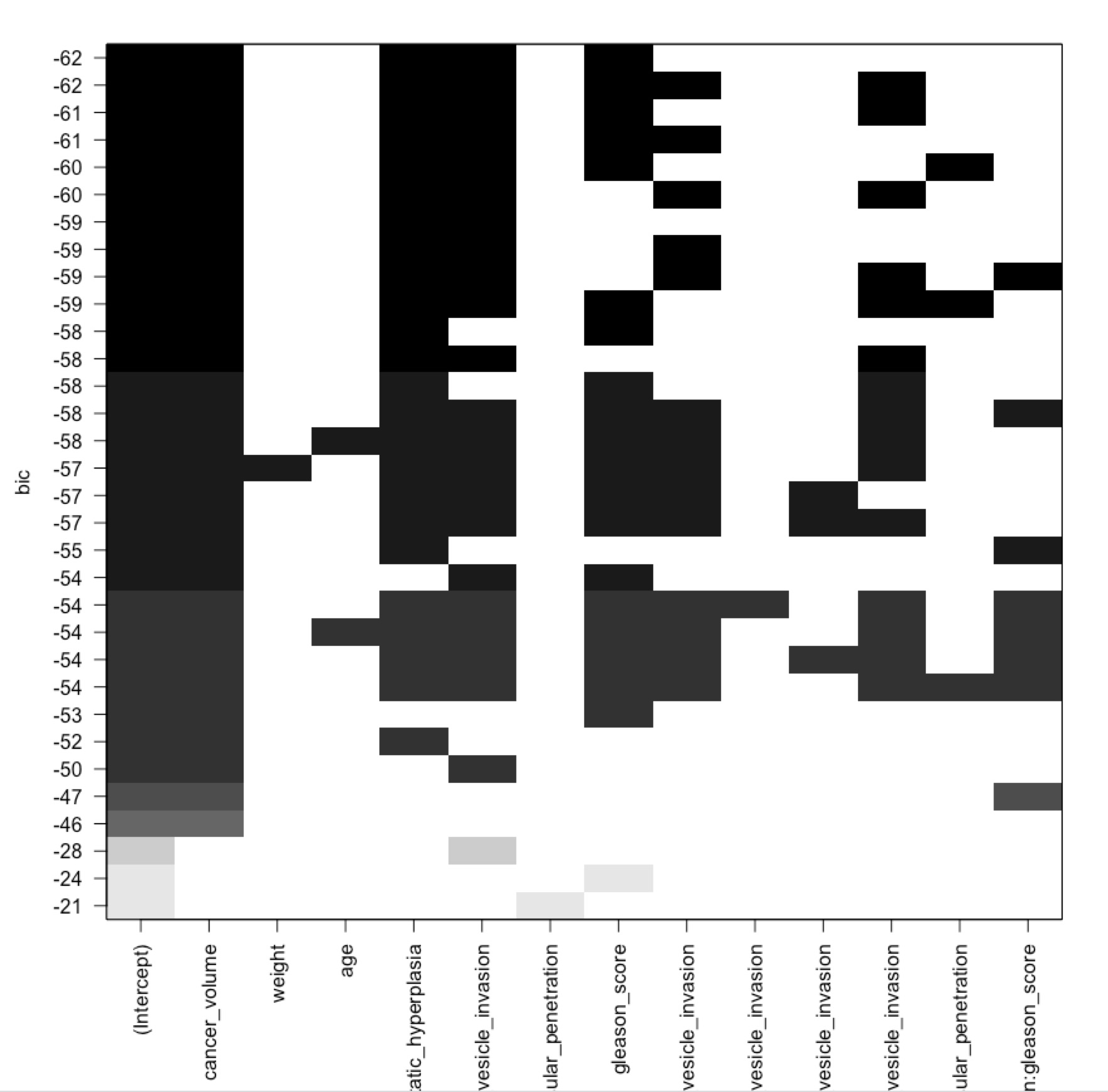


Fit the model again and check the assumption again:

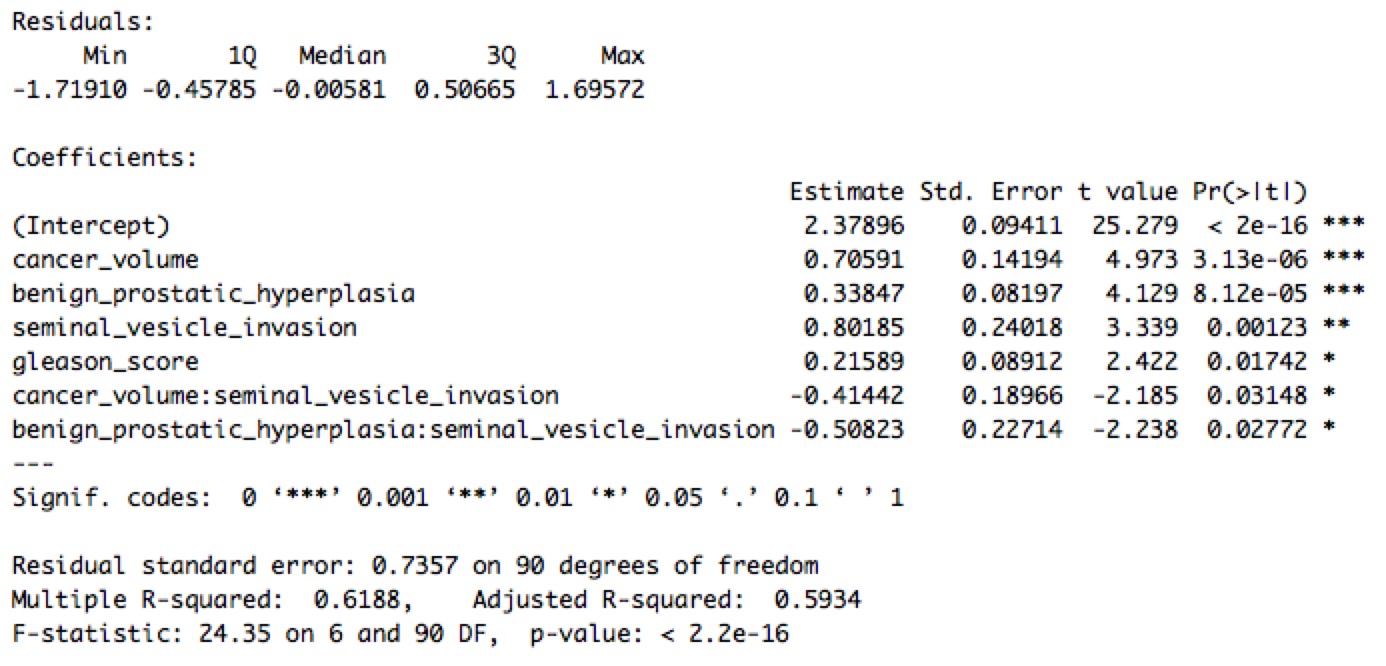


Now, we can do further analysis using the linear regression model.

Then, I tried automatic model selection:

Next, I selected the predictors that has the best adjusted R2 and relatively low BIC (low complexity), which are “cancer\_volumn”, “benign\_prostatic\_hyperplasia”, “seminal\_vesicle\_invasion”, “gleason\_score”, “cancer\_volume:seminal\_vesicle\_invasion” and “benign\_prostatic\_hyperplasia:seminal\_vesicle\_invasion”. Fit the model again and got the result:



We can see that all the selected predictors are very important. So, the final regression function is:



x\_rescaledn is the rescaled value (0 mean and 1 standard deviation) of xn

When “cancer\_volumn” (x1) is 4.2633, “weight” (x2) is 22.783, “age” (x3) is 68, “benign\_prostatic\_hyperplasia” (x4) is 1.3500, “seminal\_vesicle\_invasion” (x5) is 0, “capsular\_penetration” (x6) is 0 and “gleason\_score” (x7) is 6, the predict result log(psa\_level) would be:

(This way, we can’t get the predict interval for original response, so abandon)

Random Forest:

Recursive Feature Elimination with 3-fold cross validation and default hyperparameters:

[(1, 'age', True),

(1, 'benign\_prostatic\_hyperplasia', True),

(1, 'cancer\_volume', True),

(1, 'capsular\_penetration', True),

(1, 'seminal\_vesicle\_invasion', True),

(1, 'weight', True),

(2, 'gleason\_score', False)]

Grid search for hyperparameters:

{'max\_features': 'sqrt', 'n\_estimators': 25}